DOCUMENT RESUME

ED 468 055 TM 034 379

AUTHOR Roberts, James S.; Samuelsen, Karen
TITLE Analyzing the Structure of Binary Test Responses Using an Optimal Scaling Approach.
PUB DATE 2002-04-00
NOTE 34p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 2-4, 2002).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS *Estimation (Mathematics); *Responses; Scaling; Simulation
IDENTIFIERS *Binary Data Analysis; Binary Scores; Bootstrap Methods; Dimensionality (Tests); *Optimal Scaling

ABSTRACT

This study explored alternative methods of estimating the number of latent dimensions represented by binary test data. The alternative methods vary in their complexity and include: (1) the minimum average partial correlation technique (MAP; W. Velicer, 1976) applied to phi correlations between binary responses; (2) the bootstrapped parallel analysis method (A. Buja and N. Eyuboglu, 1992) applied to phi correlations; (3) the minimum average partial correlation technique applied to data that has been optimally scaled using a monotonic transformation in which tied data may be untied (J. Kruskal, 1964; SAS Institute, 1999); (4) the bootstrapped parallel analysis procedure applied to those same optimally scaled responses; and (5) the DETECT procedure (J. Zhang and W. Stout, 1999). The binary responses were simulated with either a one-dimensional or a two-dimensional three-parameter logistic model, and then those responses were analyzed with each of the five dimensionality estimation techniques. The mean difference between the estimated and true dimensionality suggested that the MAP and DETECT procedures performed best, on average, in conditions where the data were truly unidimensional. In contrast, the bootstrapped parallel analysis and theDETECT procedures performed best, on average, in the two-dimensional data conditions. In addition, the bootstrapped parallel analysis estimated were less sensitive to the amount of correlation between latent dimensions relative to those produced by the DETECT procedure. These limited simulations suggest that the bootstrapped parallel analysis procedure may be useful to applied researchers and students who have little access to or knowledge of technically sophisticated dimensionality assessment. (Contains 4 figures and 27 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 468 055

Analyzing the Structure of Binary Test Responses Using an Optimal Scaling Approach

James S. Roberts

Karen Samuelsen

University of Maryland

Paper presented at the annual meeting of the National Council on Measurement in Education,

New Orleans, Louisiana

TM034379

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. S. Roberts

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

Abstract

This paper explores alternative methods of estimating the number of latent dimensions represented by binary test data. The alternative methods vary in their complexity and include the minimum average partial correlation technique (Velicer, 1976) applied to phi correlations between binary responses, the bootstrapped parallel analysis method (Buja & Eyuboglu, 1992) applied to phi correlations, the minimum average partial correlation technique applied to data that has been optimally scaled using a monotonic transformation in which tied data may be untied (Kruskal, 1964; SAS Institute, 1990a), the bootstrapped parallel analysis procedure applied to those same optimally scaled responses, and the DETECT procedure (Zhang & Stout, 1999). The binary responses were simulated with either a one-dimensional or a two-dimensional, three-parameter logistic model, and then those responses were analyzed with each of the five dimensionality estimation techniques. The mean difference between the estimated and true dimensionality suggested that the MAP and DETECT procedures performed best, on average, in conditions where the data were truly unidimensional. In contrast, the bootstrapped parallel analysis and the DETECT procedures performed best, on average, in the two-dimensional data conditions. Additionally, the bootstrapped parallel analysis estimates were less sensitive to the amount of correlation between latent dimensions relative to those produced by the DETECT procedure. These limited simulations suggest that the bootstrapped parallel analysis procedure may be useful to applied researchers and students outside of the psychometrics arena who have either little access to or less knowledge about more technically sophisticated dimensionality assessment methods.

Analyzing the Structure of Binary Test Responses Using an Optimal Scaling Approach

There have been many methods proposed for dimensionality assessment of binary responses to educational test items. These include traditional principal components or factor analysis (Green, 1983; Hambleton & Rovinelli, 1986; Hambleton & Traub, 1973; Reckase, 1979), factor analysis models based on parametric item responses theory (Bock, Gibbons, & Muraki, 1988; Fraser, 1988) and models based on nonparametric item response theory (Zhang & Stout, 1999) among others. The traditional principal components (PC) approach suffers from two main difficulties. First, PC is a linear model of the response process whereas item responses to educational tests are generally thought to be a nonlinear function of a latent trait (Lord, 1980). Second, when analyzing binary responses, those items with similar difficulties may give rise to spurious "difficulty factors" (Gorsuch, 1983). McDonald & Ahlawat (1974) have argued that it is not necessarily the binary nature of the item responses that leads to the emergence of difficulty parameters, but rather the nonlinear relationship between the latent trait and the item responses.

There are several complex factor analytic models available to researchers and measurement professionals who are psychometrically savvy. These include models like those implemented in the POLYFACT (Muraki & Carlson, 1995), NOHARM (Fraser, 1988) and TESTFACT (Wilson, Wood & Gibbons, 1991) computer programs. The programs attempt to explicitly model the item responses as a nonlinear function of a multidimensional latent trait which, in turn, overcomes the two problems associated with the PC approach; namely the binary character of responses and the nonlinear relationship between responses and the latent trait. However, these models assume particular parametric relationships between the latent trait and the observed responses and are valid only to the extent that such relationships hold. Other models such as the DETECT

procedure are nonparametric in this sense, and thus, appear to be more general.

Although the complex dimensionality assessment tools mentioned above are more justifiable in a theoretical sense, they require more rigorous quantitative training for the data analyst and specialized software relative to the traditional principal components procedure. If one is introducing students and/or practitioners from other areas to the basics of testing and measurement, then more simplistic methods must be employed. The benefit of using PC-based methods to assess dimensionality arises because of their simplicity. PC is generally taught in introductory courses on multivariate data analysis or applied measurement. Moreover, most statistical computer programs perform PC analysis. Unfortunately, estimating the number of dimensions from PC-based methods may be compromised to the extent that difficulty factors are included in the estimates.

In this study, an alternative type of principal components analysis will be explored. Specifically, the responses to each item will be optimally scaled to maximize their covariance with linear combinations of the remaining test items. The optimal scaling technique is based on a method developed by Kruskal (1964) in which tied responses can be untied as long as the ordinal relationship between the original 0/1 categorization is preserved. Principal components based on the optimally scaled data results in a nonparametric representation of the underlying structure of the test items. This type of optimal scaling is available in a popular statistical computing package (i.e., the SAS PRINQUAL procedure), and the method itself can be conceptually described to individuals who possess only a modest understanding of multiple regression and principal components.

The objective of this study is to compare the classical principal components method, the

principal components of optimally scaled responses, and the DETECT procedure with regard to the ability to determine the number of dimensions inherent in a set of binary test data. These methods were chosen to represent an overly simplistic approach, a rationally based approach, and a highly sophisticated theoretical approach. Although the analysis of high-stakes testing data will always warrant more sophisticated dimensionality assessment procedures, it will be interesting to see how a ubiquitous computer program like the SAS PRINQUAL procedure can perform relative to specialty software such as DETECT which has been specifically designed for dimensionality assessment with binary test responses.

Method

Two different simulations were conducted in this study. In one simulation, binary responses were generated using a unidimensional item response model. In the second simulation, a two-dimensional item response model was used to generate responses. Each of these simulations is described in detail below.

Unidimensional Simulation

Data Generation

Data in the unidimensional simulation were generated using a unidimensional, three-parameter logistic model (Birnbaum, 1968) in which the probability of a correct response to an item was given by:

$$P_i(\theta_n) = c_i + (1 - c_i)\frac{\exp[\,a_i(\theta_n - b_i)\,]}{1 + \exp[\,a_i(\theta_n - b_i)\,]}, \qquad (1)$$

where:

θ_n is the latent ability for the *nth* simulee,

b_i is the location of the *ith* item on the latent continuum (i.e., the item difficulty),

a_i is the discrimination of the *ith* item, and

c_i is the pseudo-chance level parameter for the *ith* item.

Both test length and item difficulty were systematically varied. Five levels of test length (12, 24, 36, 48 and 60 items) were chosen to represent a range of tests from extremely short to fairly long. These test lengths were crossed with two levels of item location clustering (unclustered and clustered item locations). In the unclustered item location condition, true b_i parameters were randomly sampled from a uniform distribution along the interval (-2, +2) reflecting values often found in practice (Hambleton & Swaminathan, 1985). In the clustered item location condition, true b_i parameters were clustered into three intervals defined by (-2.0, -1.5), (-.25, +.25) and (+1.5, +2.0), and one-third of the test items were randomly located within each cluster. The clustered item difficulty condition was included to promote the occurrence of difficulty factors (Gorsuch, 1983). Item discriminations were sampled from a log normal distribution with $\mu=0$ and $\sigma=.5$, whereas pseudo-chance parameters were sampled from a beta distribution with $\alpha=5$ and $\beta=17$. These distributions of item discrimination and pseudo-chance level parameters are identical to the corresponding prior distributions used in BILOG (Mislevy & Bock, 1990). Person parameters (i.e., θ_n) for 1000 simulees were independently sampled from a standard normal distribution on each replication.

One hundred replications were performed in each experimental condition. All model parameters were independently sampled from their respective distributions on each of these

replications, and consequently, each model parameter was a random variable. It was hoped that this feature would increase the generalizability of the results. The binary response to a given item was obtained by comparing the value of a uniform (0,1) random deviate to the probability value obtained from Equation #1 using the true model parameters for a given simulee-item combination.

Dimensionality Assessment

After the 1000 response vectors were generated for a given replication, the dimensionality of the item responses was determined using each of five methods. Each method is described below.

Minimum Average Partial Correlation Criterion (MAP). The MAP procedure (Velicer, 1976) estimates the number of dimensions inherent in a set of data by finding the number of successive principal components that must be extracted in order to minimize the average squared partial correlation. The MAP criterion was applied to Pearson product moment correlation matrix associated with the binary item responses (i.e., phi coefficients).

Minimum Average Partial Correlation Criterion Based on Optimally Scaled Data (MAPOP). The MAPOP procedure was identical to the MAP procedure with the exception that principal components were calculated from item responses that were first transformed in an optimal fashion using a minimum generalized variance (MGV) method implemented in the SAS PRINQUAL procedure (SAS Institute, 1990ab). The MGV method is an iterative method in which the following steps are performed:

a) select the *ith* variable in the $N \times I$ data matrix to serve as a criterion variable in a multiple regression where the predictors are the remaining I-1 columns in the data matrix (or some full rank subset of the remaining I-1 columns)

b) obtain the predictions for the criterion variable from this regression model

c) optimally scale these predictions using a monotonic transformation (i.e., monotonic in relation to the criterion variable) that allows originally tied criterion values to become untied (Kruskal, 1964)

d) standardize the optimally scaled data to have a mean and standard deviation equal to that for the original variable

e) replace the *ith* column in the $N \times I$ data matrix with the optimally scaled data

f) repeat processes (a) through (e) for each column of the $N \times I$ data matrix

g) repeat processes (a) through (f) until the algorithm converges (i.e., until the rescaled data changes very little from one iteration to the next.)

The MGV algorithm attempts to produce a set of transformed variables that are as redundant as possible while maintaining a monotonic relationship between the transformed variables and the original variables. This type of transformation will maximize the variance produced by each successive principal component under the constraint that the transformed variable must be monotonically related to the original variables. Given the nonlinear, monotonic relationship between $P_i(\theta)$ and θ, we presumed that binary item response variables transformed in this fashion would have less tendency to yield difficulty factors in a subsequent dimensionality analysis. In that subsequent analysis, the dimensionality was estimated by applying the MAP criterion to the principal components of the transformed data.

Bootstrapped Parallel Analysis (BSP). The BSP procedure (Buja & Eyuboglu, 1992) is a resampling analog of Horn's (1965) parallel analysis procedure. In the BSP procedure, an $N \times I$ random response matrix is constructed where N is the sample size and I is the test length. Each

column of this matrix is built independently by resampling with replacement from the corresponding column in the original data matrix. In this way, the correlation among the item responses is effectively eliminated whereas the marginal distribution characteristics of responses to each item are maintained. After the random response matrix is created, the principal components of the Pearson product moment (phi) correlations derived from this matrix are calculated and the resulting eigenvalues are stored. This process is repeated t times, and the average eigenvalue obtained over these t repetitions is calculated for each successive component. These average eigenvalues are compared to the eigenvalues associated with the real responses. The estimated number of dimensions is equal to the first m components for which the eigenvalues of real data are greater than the corresponding average eigenvalues of random data. Note that, in this simulation, the creation of eigenvalues from random data was repeated t=100 times in each replication, and then eignenvalues were subsequently averaged over these 100 values.

Bootstrapped Parallel Analysis of Optimally Scaled Data (BSPOP). The BSPOP procedure is simply the BSP procedure applied to item responses that have been optimally scaled using the MGV method described above.

Dimensionality Evaluation to Enumerate Contributing Traits (DETECT). Zhang and Stout (1999) have developed a quantity known as the DETECT index to determine the number of dimensionally distinct clusters of items on a test and simultaneously assign items to those clusters. The number of dimensionally distinct clusters, K, can be used as an estimate of the number of dimensions inherent in a data set with approximate simple structure. A set of test items can be assigned to K clusters a variety of ways. Each of these different assignments constitutes a different partition, P. The theoretical DETECT index can be defined for a given partition as:

$$D(P) = \frac{2}{I(I-1)} \sum_{1 \le i < j \le I} \delta_{ij}(P)\, E\,[Cov\,(X_i, X_j \mid \Theta_T)\,] \qquad (2)$$

where:

I is the number of test items,

$E\,[Cov\,(X_i, X_j \mid \Theta_T)\,]$ is the expected conditional covariance between scores on items i and j after conditioning on the latent test composite, Θ_T,

$\delta_{ij}(P)$ is an indicator variable that is equal to +1 when items i and j are in the same cluster, and it is equal to -1 when items i and j are in different clusters.

The theoretical DETECT index is maximized when each item is assigned to the correct cluster. The DETECT procedure estimates $D(P)$ for alternative partitions, P, using a genetic algorithm. The partition that maximizes the estimate is chosen as the optimal partition of items, and the number of clusters, K, can be used as an estimate of the dimensionality of the test responses.

The DETECT procedure was implemented using commercial software also referred to as DETECT (The William Stout Institute for Measurement, 1999). The current study used the cross validation strategy recommended by Zhang and Stout (1999). Simulees were randomly divided into two halves of 500 simulees each. The DETECT index was maximized in the first half of data. Denote this index as $\hat{D}_{MAX}$. The procedure was run again in the second half of data (i.e., the cross validation data set). The optimal partition found in the cross validation data was then used to recalculate the DETECT index in the first half of data. Denote this recalculated index as $\hat{D}_{REF}$. The estimated dimensionality was set equal to 1 whenever $\hat{D}_{REF} < .1$ or $(\hat{D}_{MAX} - \hat{D}_{REF}) / \hat{D}_{REF} > .5$. The DETECT procedure conditions on both total test score and corrected test score when calculating conditional covariances between pairs of items.

Consequently, the user must specify the minimum number of respondents required at each corrected score level. This value was initially set to 20 for each replication but was then adaptively reduced on each replication in order to ensure that at least 85% of the respondents were maintained in the analysis. The user must also specify the number of mutations allowed in the genetic algorithm that is implemented in the program. This value was set to 7 for all experimental conditions.

Data Analysis

The results were analyzed using a 5 (test length) x 2 (item clustering) x 5 (dimensionality estimation method) split-plot ANOVA where the first two factors represented between-replications factors and the last was a within-replications factor. Statistical tests of effects that involved the within-replications factor were adjusted for sphericity violations using the Huynh-Feldt (1970) procedure. The dependent variable was the number of dimensions suggested by a given estimation method. One hundred replications were performed in each of the between-replications cells. The statistical significance and effect size estimate (η^2) for each ANOVA effect were examined. The η^2 values were calculated separately for between-replications and within-replications effects due to the fact that there were multiple error terms in the split-plot design. For between-replications effects, the η^2 value represented the proportion of between-replications sums of squares accounted for by the given effect. For within-replications effects, the η^2 represented the proportion of within-replications sums of squares attributable to an effect. To guard against interpreting small effects with little practical importance, only those effects that were both statistically significant ($p < .05$) and had η^2 values $\geq .02$ were deemed worthy of interpretation.

Two-dimensional Simulation

Data Generation

The second simulation was similar to the first with the primary exception that binary responses were generated with a two-dimensional, three-parameter logistic model with compensatory abilities (Reckase & McKinley, 1983) given by:

$$P_i(\underline{\theta}_n) = c_i + (1 - c_i) \frac{\exp[\sum_{k=1}^{2} (a_{ik}(\theta_{nk} - b_{ik}))] \cdot}{1 + \exp[\sum_{k=1}^{2} (a_{ik}(\theta_{nk} - b_{ik}))]}, \quad (3)$$

where:

b_{ik} is the location parameter for the *ith* item on the *kth* dimension in the latent space,

a_{ik} is the discrimination parameter for the *ith* item on the *kth* dimension in the latent space,

c_i is the pseudo-chance level parameter for the *ith* item, and

θ_{nk} is the location of the *nth* individual on the *kth* dimension in the latent space.

Note that $\underline{\theta}_n$ refers to a vector containing the *k*-dimensional coordinate locations for the *nth* individual. In this study, item location coordinates were randomly sampled from two independent uniform distributions (both again ranging from -2 to +2) or were clustered into square segments of the two-dimensional latent space. The sides of each square correspond to the cluster intervals defined in Study 1. Person locations were sampled from a bivariate standard normal distribution with a correlation equal to ρ. The value of ρ was set equal to .1 or .6 to reflect minimally or markedly oblique dimensional structures, respectively. These values are consistent with those used by Hambleton and Rovinelli (1986) in their study of dimensionality assessment. Pseudo-chance parameters were again sampled from a beta distribution as in Study 1. Item discrimination

parameters were sampled from a mixture of two lognormal distributions using a technique described by Nandakumar (1991). Specifically, a_{i1} was sampled from a lognormal distribution where the mean value of a_{i1} was equal to $(1-\Psi)\lambda$ and a standard deviation was $\sqrt{(1-\Psi)}\kappa$, whereas a_{i2} was independently sampled from a lognormal distribution where the mean value of a_{i2} was equal to $\Psi\lambda$ and a standard deviation was $\sqrt{\Psi}\kappa$. The values of λ and κ were set equal to 1.13 and .6, respectively, and these values corresponded to the hyperparameters of the lognormal prior distribution (i.e, $\mu=0$ and $\sigma=.5$) used in BILOG. (See Baker, 1992 for a description of the relationship between hyperparameters associated with the distribution of a_{i1} and those for $\log(a_{i1})$.) When a_{i1} and a_{i2} were added together, the resulting variable had a mean equal to λ and a standard deviation of κ. The Ψ parameter was set equal to 0 to produce items that were a function of only the first dimension, whereas it was set equal to 1 to produce items that represented solely the second dimension. The Ψ parameter was set to .5 to produce items that were a function of both dimensions (i.e., the complex structure condition). Together, the values of Ψ and κ determined the strength of the relationship between each dimension and the responses generated for a given item. An illustration of the prototypical types of discrimination parameters that were generated under the complex structure condition (i.e., when $\Psi=.5$) is given in Figure 1.

Insert Figure 1 About Here

The number of items that contributed to each dimension was also manipulated. In the case of simple structure, items were split either 50%/50% or 75%/25% across the two dimensions. The proportion of items assigned to each dimension was based on previous dimensionality studies (Hambleton & Rovinelli, 1986; DeChamplain & Gessaroli, 1998). In the complex structure

condition, 25% of the items reflected solely the first dimension, 25% reflected solely the second dimension, and 50% were a function of both dimensions. According to DeChamplain & Gessaroli (1998) this condition is typical of what one might encounter in practice. In summary, when considering the two item assignment strategies in the simple structure condition and the single item assignment strategy in the complex structure condition, there were a total of three structure types examined.

Dimensionality Assessment

The dimensionality of the items responses was estimated using the MAP, MAPOP, BSP, BSPOP and DETECT procedures. Each of these procedures was implemented in a fashion identical to that for the unidimensional simulation.

Data Analysis

The results from Study 2 were analyzed with a 5 (test length) x 2 (item clustering) x 3 (structure type) x 2 (correlation between dimensions) x 5 (dimensionality estimation method) factorial, split-plot ANOVA where the first four variables constituted between-replications factors and the last was a within-replications factor. There were 100 replications in each between-replications cell of the design, and the dependent variable was the number of dimensions suggested by a given method. As in Study 1, the Huynh-Feldt (1970) procedure was used to adjust the statistical significance of within-replication effects, and the statistical significance and effect size (η^2) were both used in conjunction to identify the most important effects in the ANOVA.

Results

Unidimensional Simulation

Results from the ANOVA on the number of dimensions suggested by each method in the unidimensional simulation are shown in Table 1. With regard to between-replications variability, there was a substantial effect of test length ($F(4,990)=135.93$, $MS_e = .43$, $p<.001$). The mean estimates were equal to 1.25, 1.50, 1.63, 1.78, and 1.86 as the test length increase from 12 to 60 items, respectively. Thus, the number of dimensions in the data was generally overestimated, and the degree of overestimation increased with test length. There were also some notable within-replications effects including a main effect of dimensionality estimation method ($F(4,3960)=2695.33$, $MS_e = .425$, $p_{adjusted}<.001$) and an interaction of dimensionality estimation method and test length ($F(16,3960)=97.30$, $MS_e = .425$, $p_{adjusted}<.001$). The average dimensionality estimate was equal to .88 for MAP, 1.05 for MAPOP, 1.15 for DETECT, 1.47 for BSP, and 3.48 for BSPOP. Thus, the MAP method tended to underestimate the dimensionality of the data, but this tendency was counteracted when the data were optimally scaled. In contrast, the BSP procedure tended to overestimate the dimensionality of the responses, and this tendency was enhanced when this method was applied to data that were optimally scaled. The DETECT procedure also tended to overestimate the dimensionality of the data, albeit to a smaller degree than BSP.

Insert Table 1 and Figure 2 About Here

The mean dimensionality estimates associated with the interaction between estimation method and test length are shown in Figure 2. With a small test length of 12 items, the MAPOP and BSP

procedures produced the most accurate mean dimensionality estimates. The DETECT procedure, in contrast, overestimated the dimensionality of the data as did the BSPOP method. The MAP procedure underestimated the dimensionality. For larger tests of 24 to 60 items, the DETECT, MAP and MAPOP estimates were all relatively accurate on average, whereas the BSP tended to overestimate the dimensionality. The overestimation observed with the BSP became more pronounced as the test length increased. The BSPOP method produced an even greater degree of overestimation, and again, the degree of overestimation increased with test length.

Two-dimensional Simulation

Table 2 portrays the statistical significance and effect size estimate for each ANOVA effect explored in the two-dimensional simulation. There were three notable between-replications effects observed when binary responses were simulated from the two-dimensional model. There was an effect of test length in which the average estimated dimensionality increased as the length of the test grew larger ($F(4,5940)=1340.46$, $MS_e=.750$, $p<.001$). The mean estimate was equal to 1.64, 2.02, 2.31, 2.51, and 2.67 as the test length increased from 12 to 60 items. The structure type also affected the mean number of dimensions estimated ($F(2,5940)=231.45$, $MS_e=.750$, $p<.001$), albeit, only slightly. The mean number of dimensions found in the case where half of the items were assigned to each factor was equal to 2.38. In contrast, the mean was equal 2.12 when the items were assigned to factors in a 1 to 3 ratio. The mean number of items was similar (i.e., 2.20) in the case where one-fourth of the items were assigned to each factor while the remaining items each represented both factors to some degree. The effects of increasing the correlation between the two dimensions was also noticeable ($F(1,5940)=1516.28$, $MS_e=.750$, $p<.001$). The number of estimated dimensions was larger when the correlation between dimensions was small

relative to when the correlation was relatively large (2.43 versus 2.04).

Insert Table 2 About Here

There were also three noteworthy within-replications effects. The effect of dimensionality estimation method was substantial ($F(4, 23760)=12250.67$, $MS_e=.762$, $p_{adjusted}<.001$). The average number of dimensions estimated with each method was equal to 1.08 for MAP, 1.41 for MAPOP, 2.18 for DETECT, 2.19 for BSP, and 4.29 for BSPOP. Thus, the MAP procedure underestimated the number of dimensions, but this underestimation was mitigated somewhat by the optimal scaling procedure. In contrast, estimates derived from the DETECT and BSP procedures were both relatively more accurate. However, when optimal scaling was combined with the BSP, the number of dimensions was substantially overestimated. There was also an interaction between estimation method and test lengtn ($F(16, 23760)=363.90$, $MS_e=.762$, $p_{adjusted}<.001$). The mean estimates corresponding to this interaction are portrayed in Figure 3. For tests of 12 items, the BSP procedure yielded an accurate estimate of the actual number of dimensions. In contrast, the MAP and MAPOP procedures severely underestimated the true dimensionality whereas the DETECT and BSPOP methods overestimated the dimensionality. For tests of 24 to 60 items the DETECT procedure provided the most accurate average estimates followed by the BSP method. The BSPOP method consistently overestimated the true number of dimensions, and this overestimation increased with test length. In contrast, both the MAP and MAPOP consistently underestimated the true number of dimensions, but the degree of underestimation attenuated somewhat as the test length increased. There was also an interaction between estimation method and the degree of correlation between dimensions ($F(4,23760)=438.22$, MS_e

=.762, $p_{adjusted}$<.001). The means corresponding to this interaction are shown in Figure 4. The mean number of estimated dimensions decreased for all methods when the correlation between dimensions increased. However, the decrease observed with the DETECT method was noticeably larger than for the other methods. As seen in the figure, the average number of dimensions suggested by the BSP method was the most accurate when portrayed as a function of the correlation between dimensions.

Insert Figures 3 and 4 About Here

Discussion

An unfortunate finding that emerged from these simulation results was that the optimal scaling transformation did not produce the desired effect. It was hoped that the transformation would help increase the linear relationship between the item responses and the underlying dimensions. This, in turn, should have logically reduced the estimated number of dimensions derived with either MAPOP or BSPOP relative to MAP and BSP, respectively. Both the MAP and BSP procedures relied on decomposing a matrix of phi correlations and were presumably susceptible to difficulty factors. However, the results from the simulations exhibited a pattern that was unexpected. When optimal scaling was applied in conjunction with traditional MAP or BSP strategies, the number of estimated dimensions increased on average. A *post hoc* examination of the optimally scaled data and associated eigenvalues showed that the technique often increased the eigenvalues for the first few components beyond those corresponding to the real latent traits. This suggests that the optimal transformation used here actually increased the impact of difficulty factors on the estimated number of dimensions. It is possible that other optimal scaling methods

might not suffer from this problem. For example, the maximum total variance method (Young, Takane, and de Leeuw, 1978) used in the SAS PRINQUAL procedure allows the user to optimally rescale data so that only the first *M* eigenvalues of a correlation matrix are maximized. However, specifying the number of eigenvalues to maximize when trying to determine the optimum number of dimensions seems logically circular, and thus, the maximum total variance method of optimal scaling was not pursued in this study.

Another interesting finding in these results is that the MAP procedure generally estimates the number of dimensions quite well when the latent trait is unidimensional. However, the procedure substantially underestimates the number of dimensions when the latent trait is two-dimensional, although the degree of underestimation attenuates as the number of test items grows larger. The tendency for the MAP procedure to underestimate dimensionality was demonstrated clearly by Zwick and Velicer (1986) in the case of continuous variables that were linearly related to latent traits in a factor analysis model. However, Zwick and Velicer downplayed the practical significance of this finding by suggesting that the MAP procedure underestimated dimensionality because it generally did not identify the most "poorly defined components". The authors even implied that this might be an advantage for the MAP procedure. In contrast, the results of this limited simulation suggest that the underestimation incurred with the MAP procedure is sometimes substantial even when factors are reasonably well defined.

Most measurement theories assume that the dimensionality of test items is either implicitly or explicitly known. In most of these theories, overestimation of the number of latent dimensions represented in test responses will simply make measurement tasks more cumbersome. For example, notions of validity must be more complex, multidimensional measurement models must

be used instead of simpler unidimensional models, or the responses to test items must be split into several essentially unidimensional subtests which are subsequently analyzed separately. In contrast, underestimation of the number of latent dimensions represented by test items can lead to more severe consequences. When the dimensionality of test items is underestimated, test score validity is suspect, and issues such as test bias arise. In item response models, the underestimation of dimensionality will generally lead to local dependence of item responses, and thus, even the most basic notions about the form of the likelihood function are disturbed. Thus, the consequences seem to be less severe when dimensionality is overestimated rather than underestimated. The value of the MAP procedure in dimensionality assessments of binary item responses seems questionable given this logic and the results of these limited simulations.

The DETECT procedure was chosen to represent a "state of the art" means to identify the dimensionality of binary test items. Moreover, the procedure performed quite well on average in both the undimensional and two-dimensional simulations. The largest average discrepancies in dimensionality estimates produced with the DETECT procedure occurred with small tests of 12 items, but these quickly dissipated with larger tests. However, the DETECT procedure was sensitive to the amount of correlation between the latent dimensions in the two-dimensional case. It exhibited a substantial degree of dimensionality overestimation when the correlation between the latent dimensions was .1, and it underestimated the number of dimensions noticeably when this correlation was equal to .6. The two types of discrepancies tended to cancel out overall. This behavior was not too surprising because the DETECT procedure was designed to perform best when approximate simple structure holds in the context of orthogonal dimensions. Nonetheless, it appears that even a method as theoretically and technically elegant as the DETECT procedure is

not without its problems.

Given the complexity of the DETECT algorithm and corresponding theory, it is unlikely that truly applied measurement practitioners or students outside of the psychometrics area would generally use the procedure. Moreover, the DETECT procedure can only be implemented with specialized commercial software. The fact that it cannot be calculated with commonly available statistical analysis software will, no doubt, constitute an additional barrier for this segment of individuals.

The BSP procedure was not without its faults. It generally overestimated the number of dimensions in the unidimensional case, and this overestimation became substantial as the test length increased. However, it performed relatively well in the two-dimensional simulation regardless of test length. The small amount of estimation error that occurred with the BSP procedure in the two-dimensional case was typically due to overestimation. Given the aforementioned logical argument about overestimation being the lesser of two evils, these results suggest that BSP procedure would be a better choice than the MAP procedure when estimating the dimensionality of binary test responses. The BSP procedure was also fairly robust to changes in the correlation between latent dimensions. In this respect, the BSP method performed better than the DETECT procedure.

The BSP is preferable to its original parallel analysis counterpart (Horn, 1965) because it maintains the characteristics of the marginal distribution of responses to each item. In contrast, applications of traditional parallel analysis typically use some arbitrary distribution from which random responses are sampled (e.g., the standard normal distribution). This arbitrary distribution may not adequately represent the distributional characteristics of real item responses, and in such

cases, the results of parallel analysis are suspect. The BSP procedure is relatively simple from a conceptual standpoint, and it can be easily implemented with publicly available SAS or SPSS macro programs. Thus, there are few, if any, barriers to using this method. When these characteristics are combined with the results of the simulations described above, the BSP procedure has much to offer in applied measurement tasks where the dimensionality of binary test items is in question.

Limitations

As is the case with any simulation study, the results are limited in scope and may not reflect the wide variety of binary item characteristics that occur in practice. Therefore, these results should be viewed as preliminary in nature. Another limitation of this work is that it has investigated dimensionality assessment only from the perspective of a naive user who simply wants to know how many dimensions exist in a set of binary test responses. This study made no attempt to ascertain which items corresponded to each dimension and if such correspondence mimicked the true latent structure. Similarly, there was no attempt to determine which, if any, dimensions were so poorly determined that they should be ignored. These are interesting questions which might lead one to make alternative judgments about the procedures evaluated in this study.

Conclusions

Determining the dimensionality of binary test responses is not an easy task. All of the dimensionality assessment procedures examined in this study showed both strengths and weaknesses. The bootstrapped parallel analysis procedure (BSP) performed reasonably well in

the two-dimensional case, although it substantially overestimated the number of dimensions in the unidimensional case. If overestimation can be tolerated more than underestimation, then the BSP procedure seems better suited to estimating the dimensionality of binary test items than does the minimum average partial (MAP) method. The results from the BSP method in the two-dimensional case were comparable to those from the DETECT procedure on average. Moreover, the BSP procedure was more robust to the level of correlation between latent dimensions than the DETECT procedure. Given its reasonable performance, its conceptual simplicity and its availability from within common statistical computing programs, the BSP procedure may prove useful to applied practitioners and students who need a relatively easy way to estimate the dimensionality of binary test responses.

References

Baker, F. B. (1992). *Item response theory: Parameter estimation-techniques.* New York: Dekker.

Birnbaum, A. (1968). Some latent trait models and their use in inferrin an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research, 27,* 509-540.

Fraser, C. (1988). *NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.* New South Wales, Australia: Center for Behavioral Studies, University of New England.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Green, S. B. (1983). Identifiability of spurious factors using linear factor analysis with binary items. *Applied Psychological Measurement, 7,* 139-147.

Hambleton, R. K., & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10,* 287-302.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Hambleton, R. K., & Traub, R. E. (1973). Analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology, 26,* 195-211.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179-185.

Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurement designs have exact F-distributions. *Journal of the American Statistical Association, 65,* 1582-1589.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika, 29,* 115-129.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27,* 82-99.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models* (2nd ed.). Chicago: Scientific Software International.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19,* 73-90.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28,* 99-117.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207-230.

Reckase, M. D., & McKinley, R. L. (1983, April). *The definition of difficulty and discrimination for multidimensional item response theory models.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

SAS Institute (1990a). *SAS/STAT user's guide, Version 6* (4th ed., vol. 2). Cary, NC: SAS Institute.

SAS Institute (1990b). *Algorithms for the PRINQUAL and TRANSREG procedures.* SAS technical report R-108. Cary, NC: SAS Institute.

The William Stout Institute of Measurement (1999). *DETECT: IRT-based educational and psychological measurement software.* Champaign, IL: The William Stout Institute of Measurement.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41,* 321-327.

Wilson, D., Wood, R., & Gibbons, R. D. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Mooresville, IN: Scientific Software.

Young, F. W., Takane, Y., & de Leeuw, J. (1978). The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika,* 43, 279-281.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 213-249.

Table 1. ANOVA results from simulations of one-dimensional item responses. Effects that are both statistically significant and account for at least two percent of the between-replications or within-replications sums of squares are denoted with bold font. Note: I=test length, C=item cluster condition, M=dimensionality assessment method.

Source	p-value	η^2
I	**<.001**	**0.352**
C	0.002	0.006
I*C	0.687	0.001
M	**<.001**	**0.659**
M*I	**<.001**	**0.095**
M*C	0.003	0.001
M*I*C	0.016	0.002

Table 2. ANOVA results from simulations of two-dimensional item responses. Effects that are both statistically significant and account for at least two percent of the between-replications or within-replications sums of squares are denoted with bold font. Note: I=test length, C=item cluster condition, S=structure type, R=correlation between dimensions M=dimensionality assessment method.

Source	p-value	η^2
I	**<.001**	**0.387**
C	0.001	0.001
I*C	<.001	0.003
S	**<.001**	**0.033**
I*S	0.059	0.001
C*S	<.001	0.004
I*C*S	0.046	0.001
R	**<.001**	**0.109**
I*R	<.001	0.006
C*R	0.857	0.000
I*C*R	0.013	0.001
S*R	<.001	0.010
I*S*R	<.001	0.006
C*S*R	<.001	0.007
I*C*S*R	<.001	0.002

Table 2. continued.

Source	p-value	η^2
M	**<.001**	**0.589**
M*I	**<.001**	**0.070**
M*C	<.001	0.001
M*I*C	<.001	0.002
M*S	<.001	0.012
M*I*S	<.001	0.003
M*C*S	<.001	0.001
M*I*C*S	0.380	0.000
M*R	**<.001**	**0.021**
M*I*R	<.001	0.007
M*C*R	0.062	0.000
M*I*C*R	0.447	0.000
M*S*R	<.001	0.001
M*I*S*R	<.001	0.003
M*C*S*R	<.001	0.003
M*I*C*S*R	0.001	0.001

Figure 1. Prototypical values of discrimination parameters for the two-dimensional, complex structure case with 60 items.

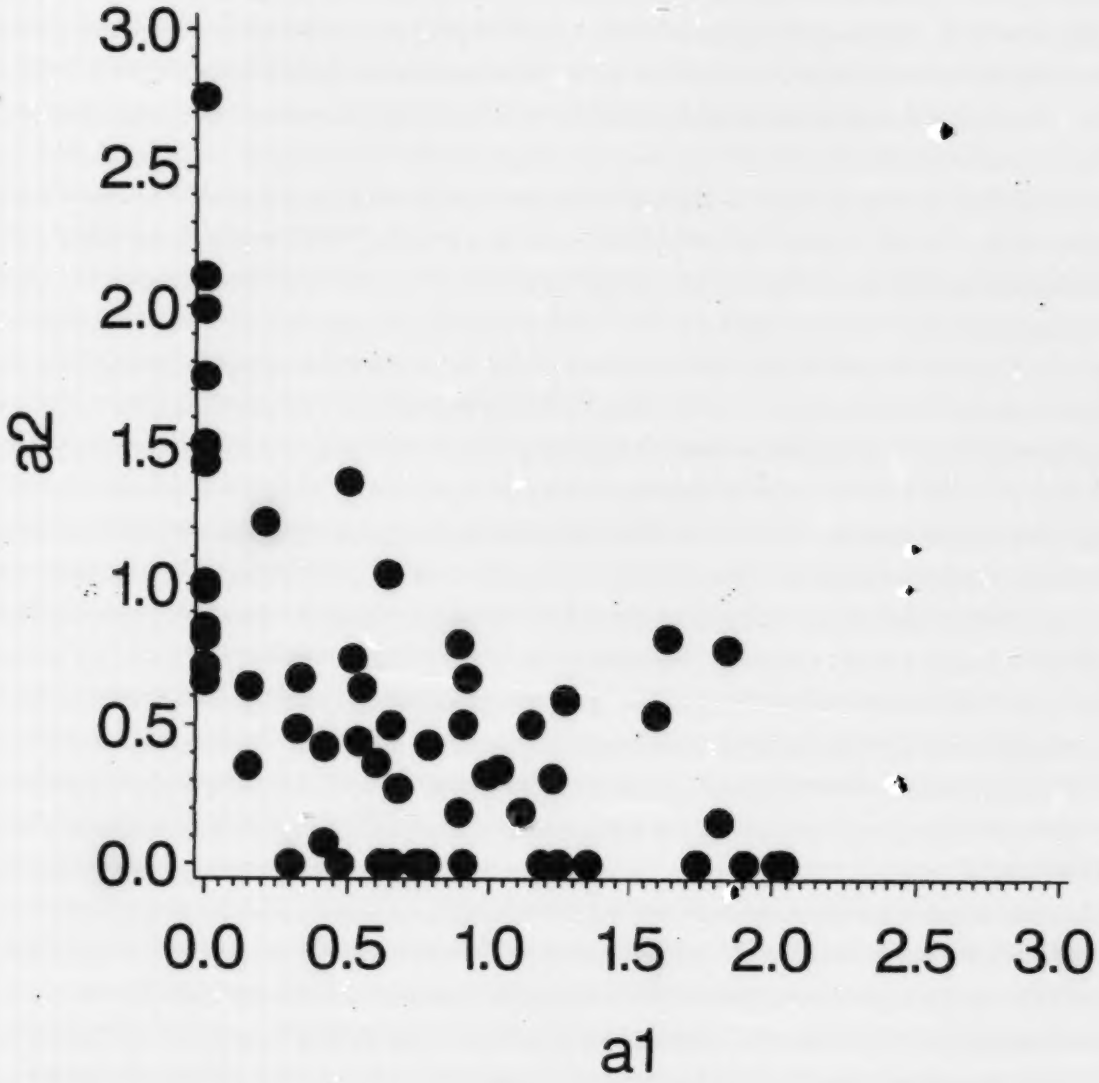


Figure 2. Average number of estimated dimensions in the unidimensional scenario by estimation method and test length.

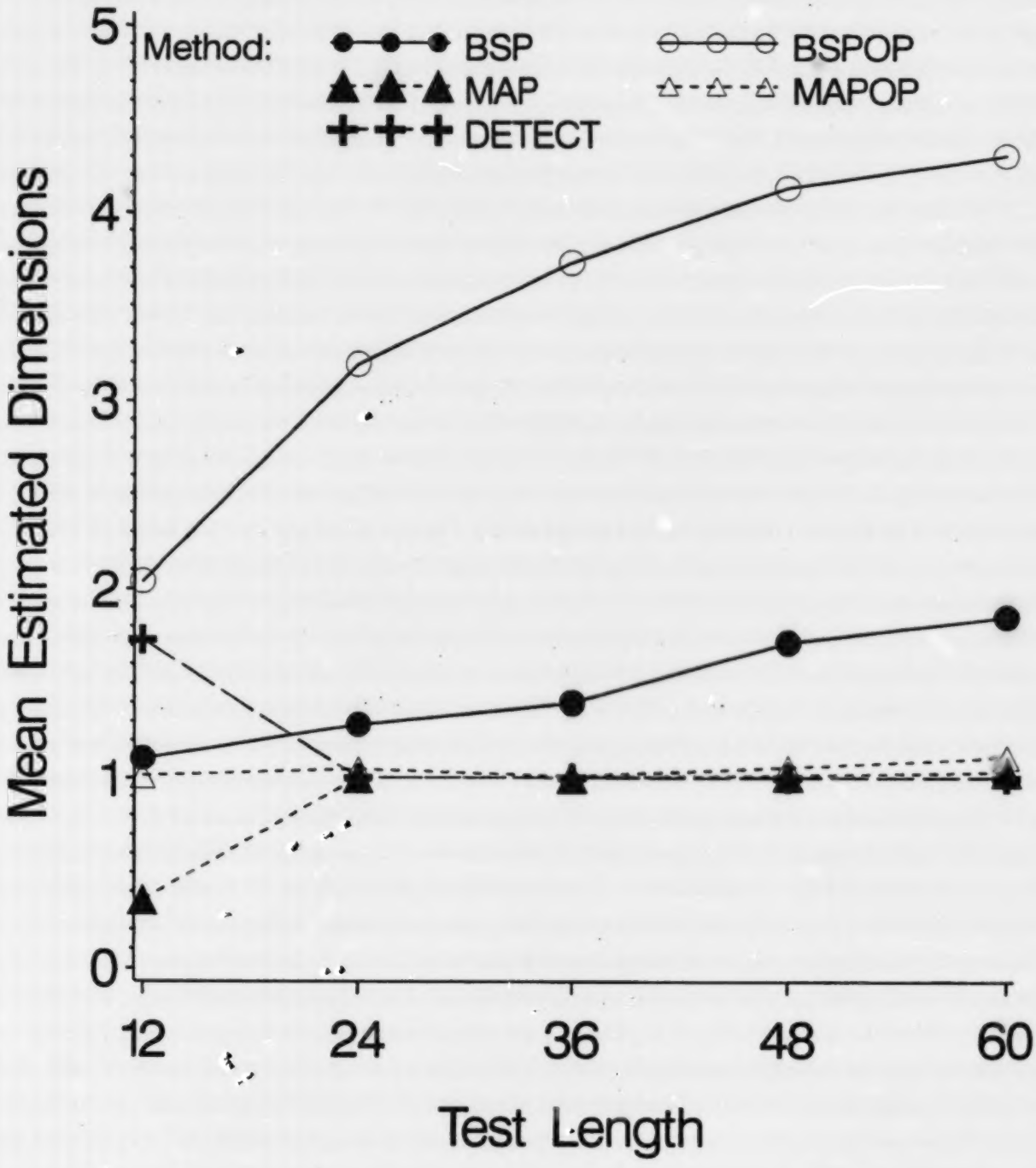


Figure 3. Average number of estimated dimensions in the two-dimensional scenario by estimation method and test length.

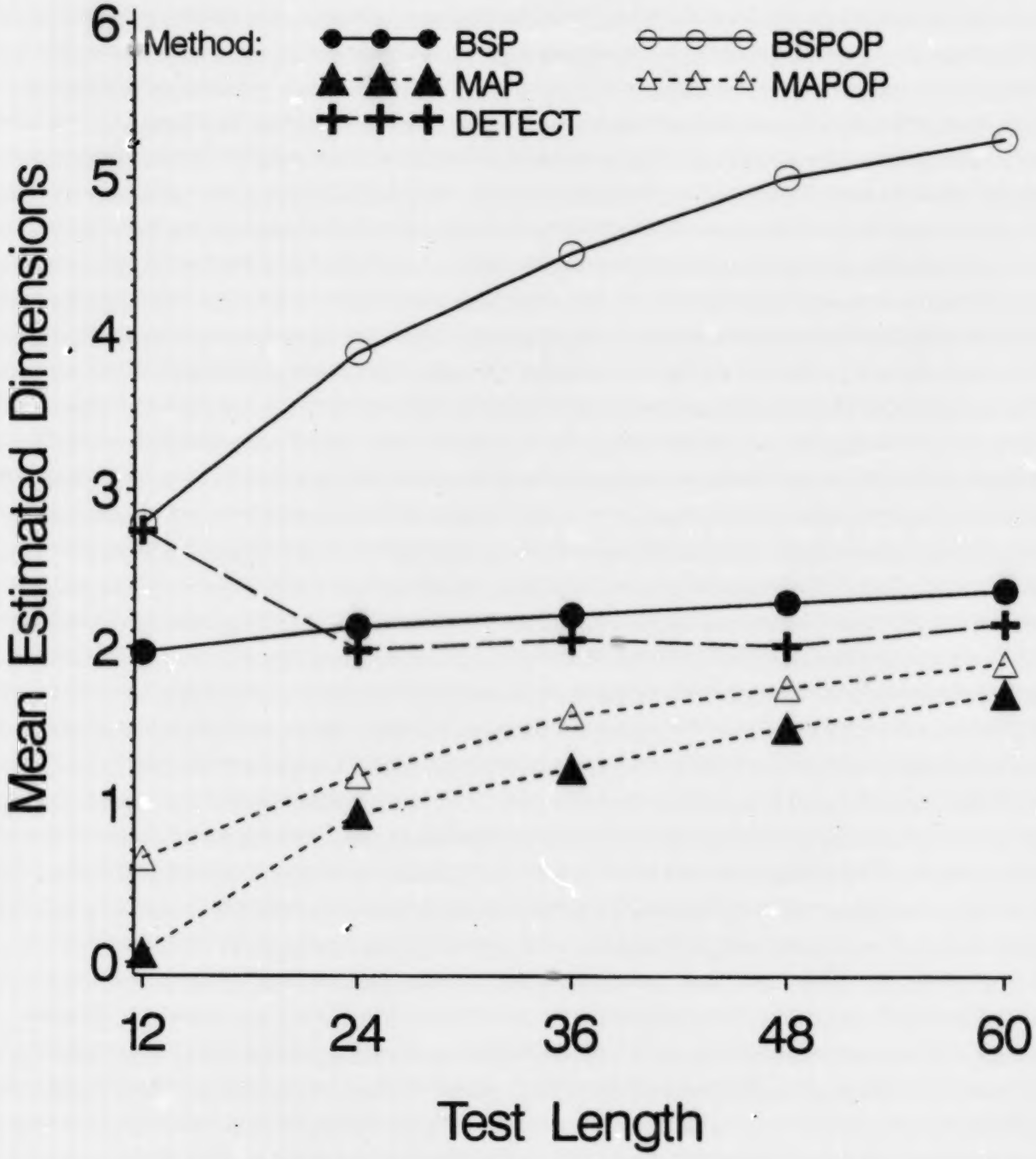
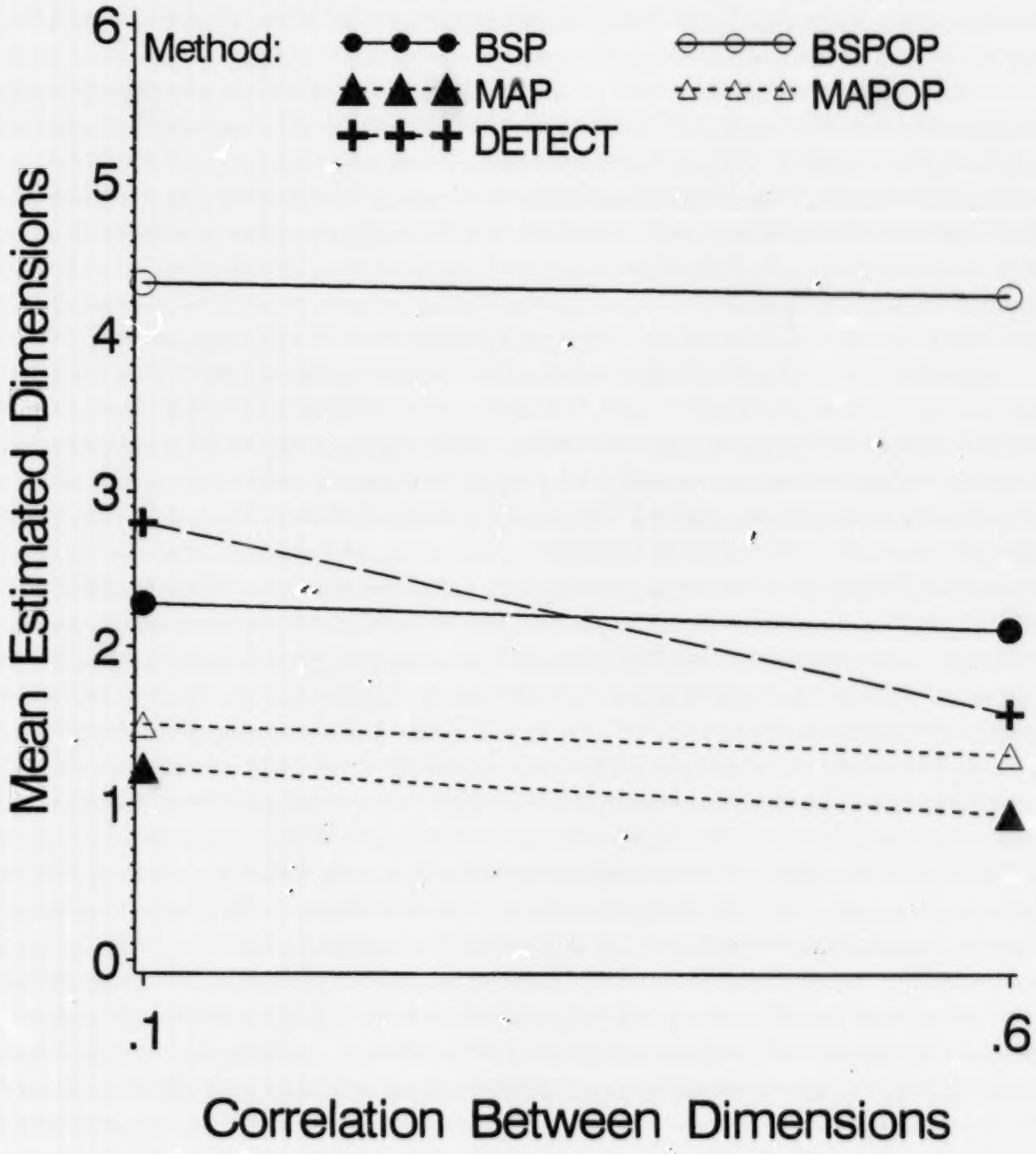


Figure 4. Average number of estimated dimensions in the two-dimensional scenario by estimation method and degree of correlation between dimensions.



U.S. Department of Education
Office of Educational Research and Improvement (OERI)

National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

TM034379

Reproduction Release

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Analyzing the Structure of Binary Test Responses Using an Optimal Scaling Approach

Author(s): James S. Roberts & Karen Samuelson

Corporate Source: | Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed t Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN ROFICHE, AND IN ELECTRONIC MEDIA ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	**Level 2A**	**Level 2B**
☑	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche o

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: James S. Roberts	Printed Name/Position/Title: James S. Roberts / Assistant Professor
Organization/Address: University of Maryland Dept. of Measurement, Statistics & Evaluation. 1230F Benjamin Bldg. College Park, MD 20742	Telephone: (301) 405-3630
	Fax: (301) 314-9245
	E-mail Address: jr245@umail.umd.edu
	Date: 7/18/02

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM: